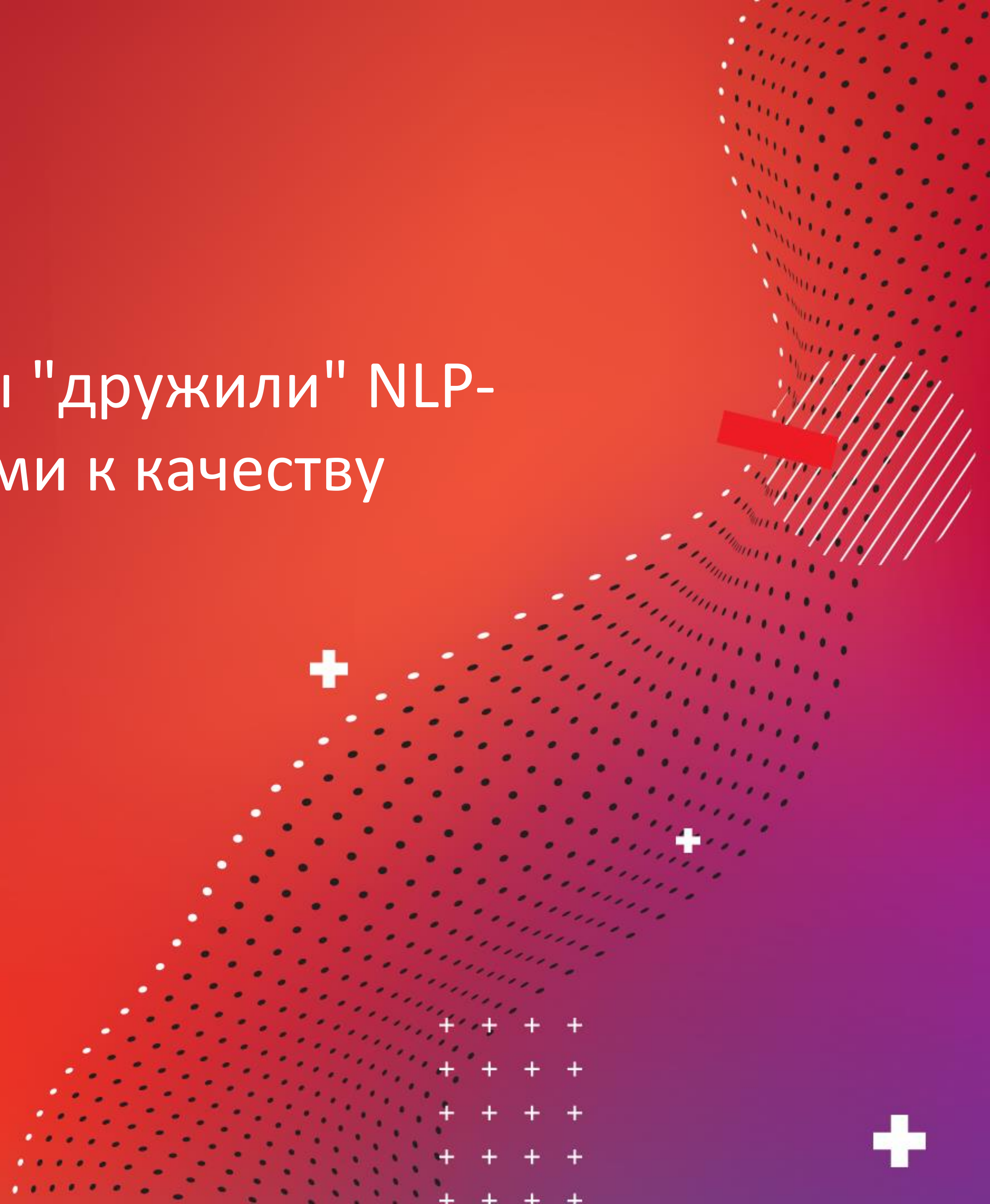


Когда трансформеры врут: как мы "дружили" NLP-  
решения с высокими требованиями к качеству

Артем Бондарь, SamsungNEXT



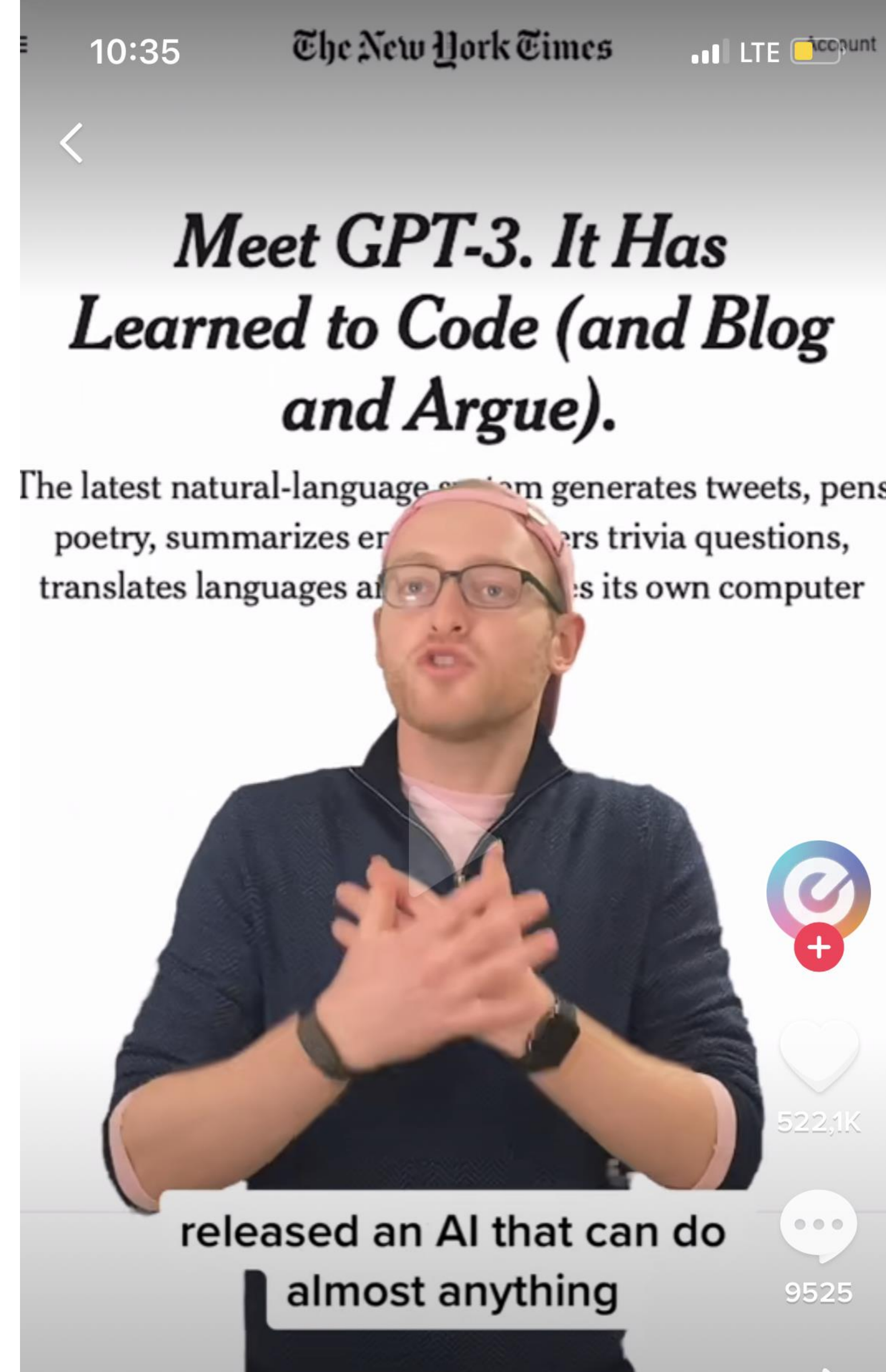
**HighLoad++**  
Весна 2021





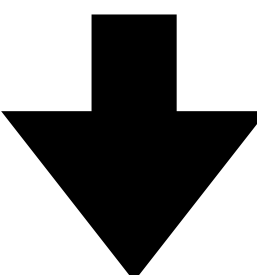
# GPT-3 hype

“It will literally solve all the problems. All you have to do is to give a clear task definition”



В это время наши  
продукты:

Qty		Rng	Unit	Prod		Comm			
3	-	4	cups	apples	honey	crisp	or	granny	smith apples ,



Qty		Rng	Unit	Brand	Prod		Comm		
3	-	4	cups	apples	honey	crisp	or	granny	smith apples ,

Почему это бренд?

# “Это норма”:

Attack text label iPod ▾



Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%



Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%

When we put a label saying “iPod” on this Granny Smith apple, the model erroneously classifies it as an iPod in the zero-shot setting.



# Пропасть между SoTA DeepLearning и продакшном



# О чем доклад:

Фреймворк, в котором мы размышляли, проектируя систему

Ошибки, за которые мы дорого заплатили

# Пару слов о себе

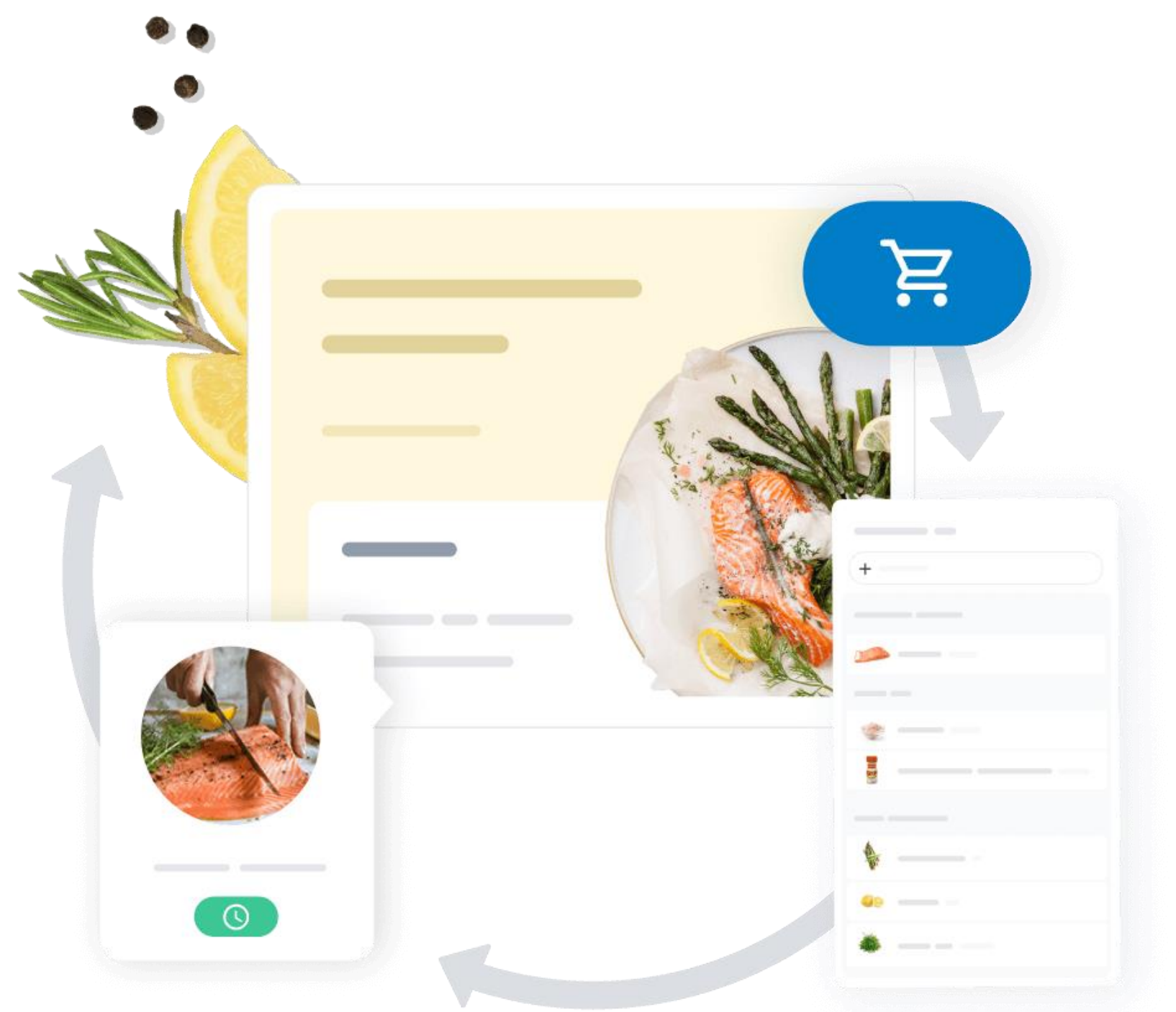


- МФТИ (ФУПМ)
- Engineering in Parallels
- Tech lead в нескольких стартапах
- MachineLearning team lead в SamsungNEXT

# SAMSUNG NEXT







# Whisk

Walmart 

ASDA

TESCO

Whisk

Kraft

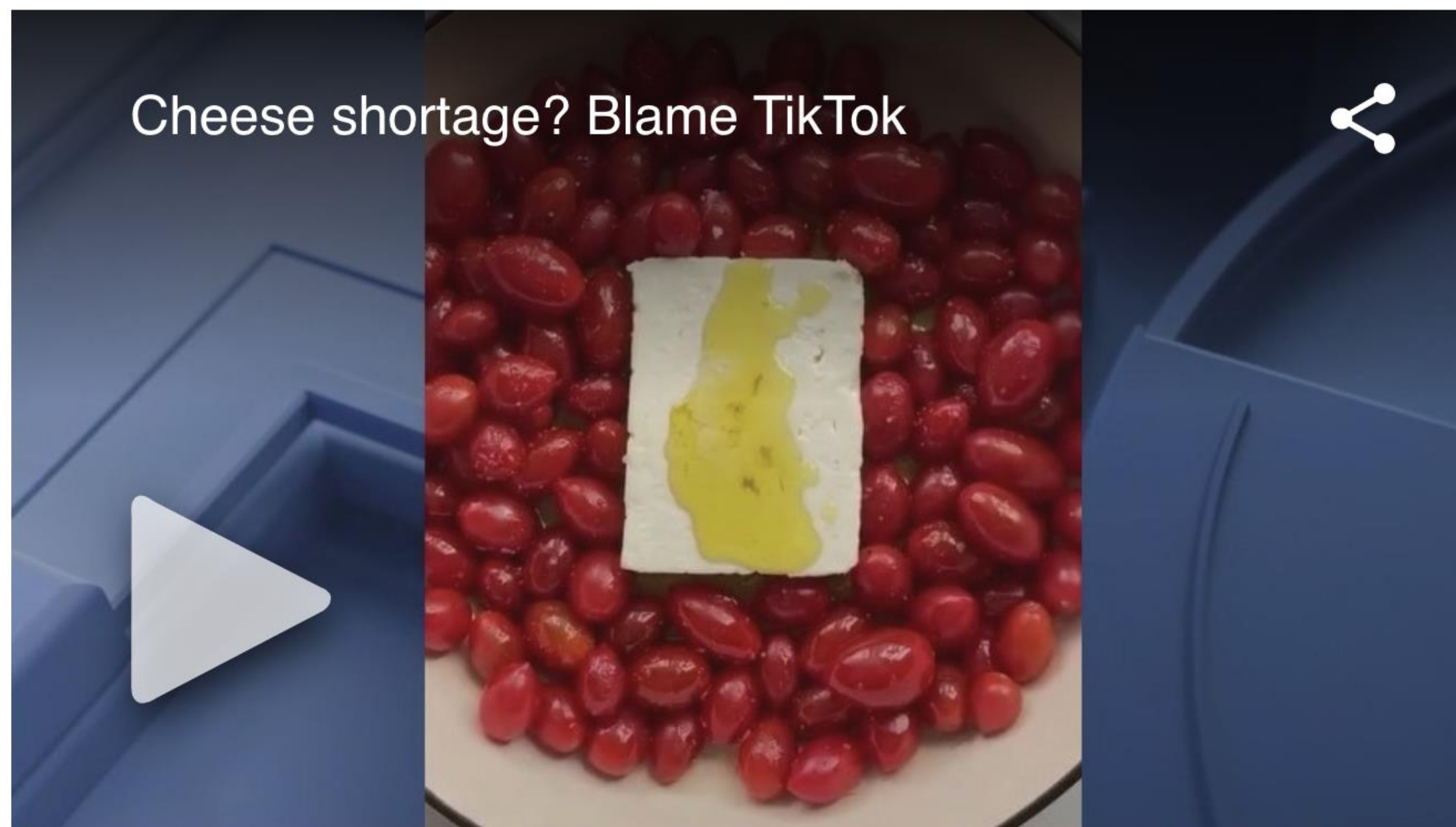






# Viral TikTok video recipe prompts feta cheese shortage

By Mac King | Published March 4 | Food and Drink | FOX 5 NY



## Cheese shortage? Blame TikTok

A TikTok video of a pasta dish based on a three-year-old blog post recipe has suddenly caused a shortage of feta cheese.

**NEW YORK** - It all started with a [pasta dish](#).

## Latest News

[View More](#)

Chicken wing shortage: Restaurants nationwide worried over skyrocketing prices



Man fraudulently obtained federal coronavirus relief funds to buy alpaca farm, prosecutors say



COVID-19 vaccine boosters likely needed every 9 to 12 months, Moderna president says



6th-grade girl shoots 3 at Idaho school before being disarmed by teacher, authorities say



Twitter introduces Tip Jar, allowing users to send money to others





# CES 2021 Презентация концепта “Smart Kitchen”

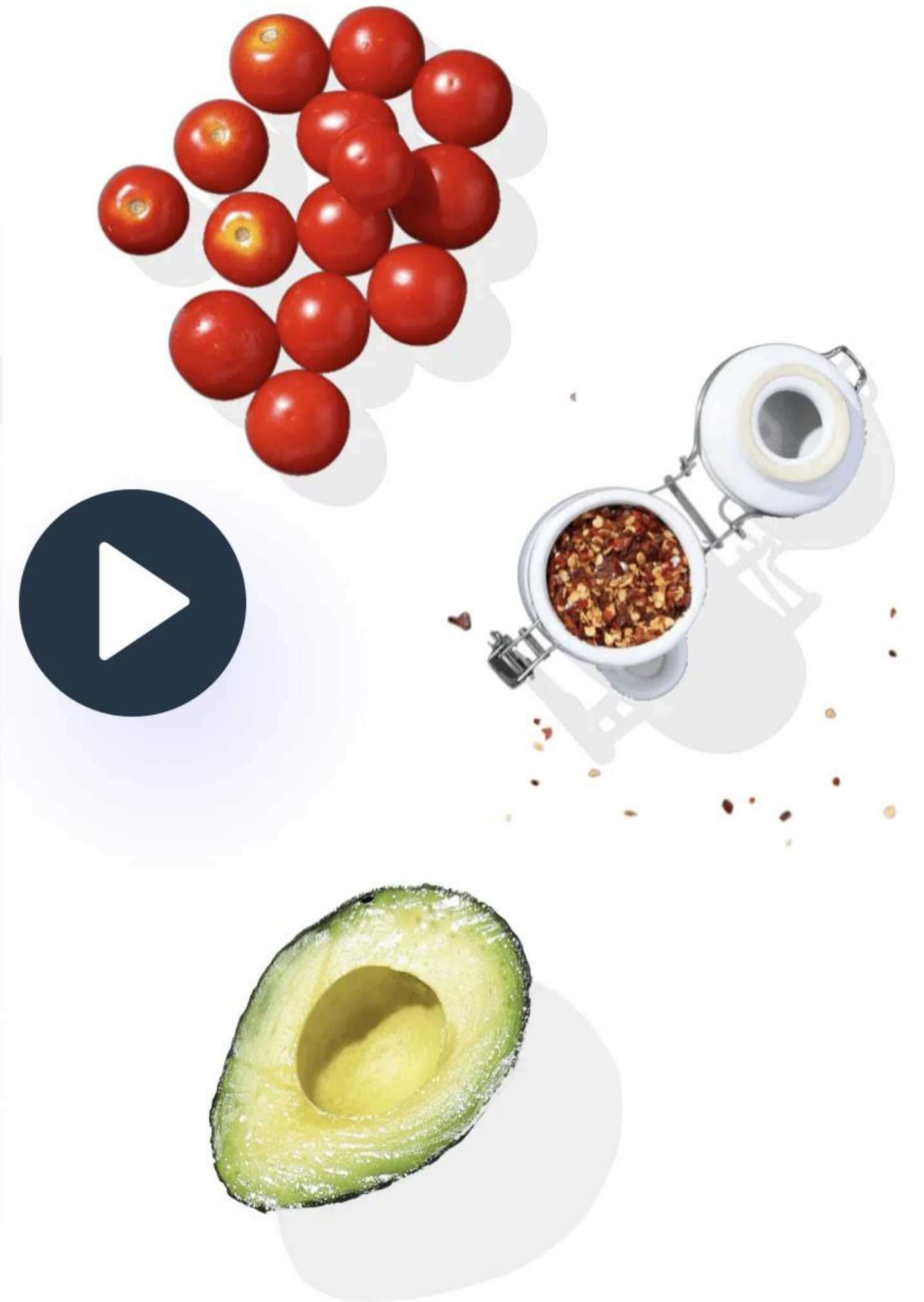
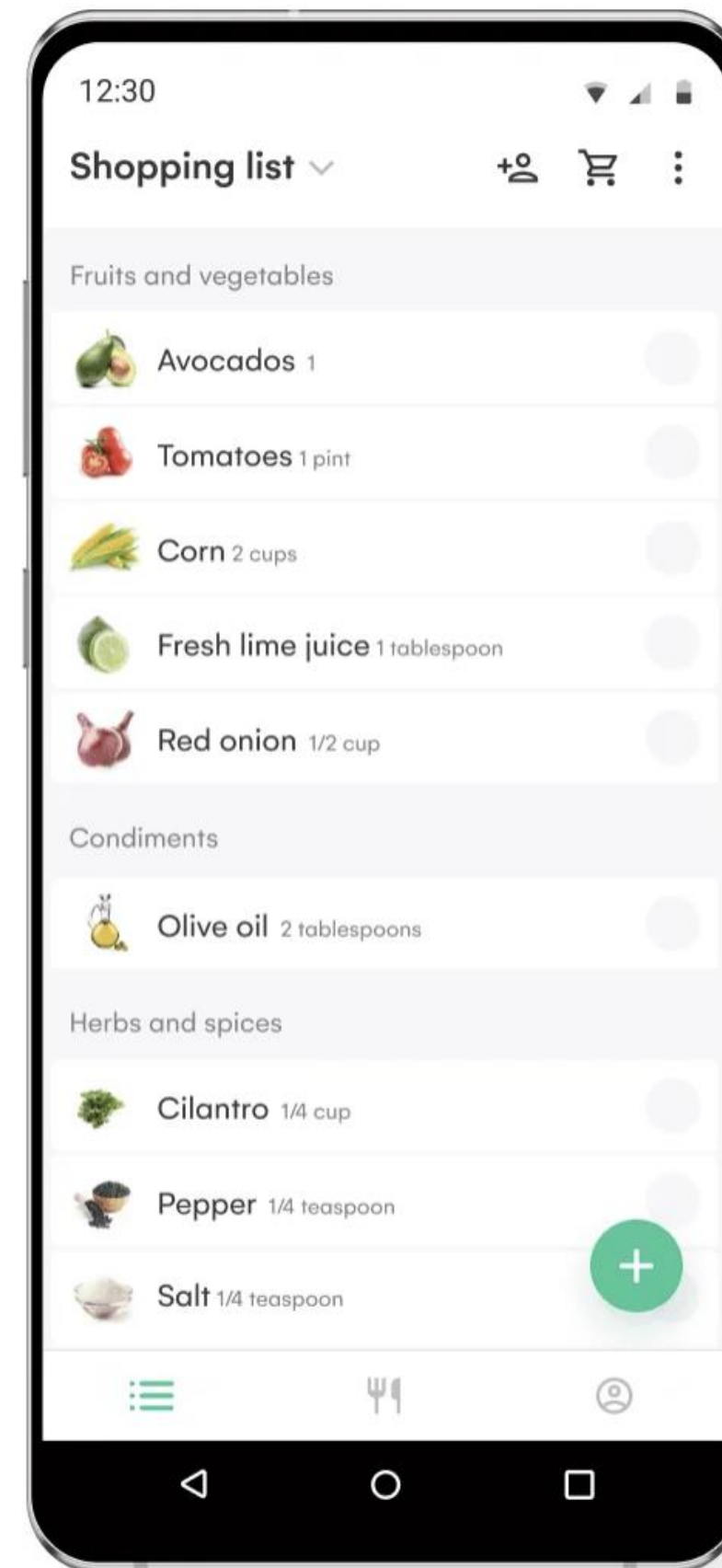


# Turn recipes into shopping lists.

Save, organize, shop, and share with Whisk. It's cooking, connected.

**Sign Up**

**Download the App**





# MachineLearning в Whisk

Категоризация рецептов

Рекомендательные системы

Умный поиск

ML

Рецепт -> список покупок

Распознавание списков покупок

Подсчет нутришн-информации

Нормализация данных для аналитики



# Как превратить рецепт в список покупок



- 100g plain flour
- 2 large eggs
- 300ml milk
- 1 tbsp sunflower or vegetable oil, plus a little extra for frying
- lemon wedges to serve (optional)
- caster sugar to serve (optional)

# Как превратить рецепт в список покупок



- 100g plain flour
- 2 large eggs
- 300ml milk
- 1 tbsp sunflower or vegetable oil, plus a little extra for frying
- lemon wedges to serve (optional)
- caster sugar to serve (optional)



**Allinson Strong White  
Bread Flour 1.5Kg**



**Tesco British Semi  
Skimmed Milk 2.272L, 4  
Pints**



**Tesco 6 Eggs**



**Tesco Olive Oil 1L**



**Tesco Lemons 4 Pack**



**Tesco Golden Caster 1Kg  
Bag**

Люди едят ингредиенты очень по-разному

2 (15 ounce) cans black beans, rinsed and drained



# Люди пишут ингредиенты очень по-разному

Так сколько вешать в граммах?

2 (15 ounce) cans black beans, rinsed and drained

Не просто фасоль, а фасоль в банке

А это, вообще, часть инструкций к приготовлению

Люди едят ингредиенты очень по-разному

3 tbsp ready-toasted mixed seeds, such as sunflower, pumpkin and  
sesame (from major supermarkets and health food shops)

Война и мир...

Люди едят ингредиенты очень по-разному

cucumber and/or cocktail shrimp for garnish (optional)

Два альтернативных ингредиента в строке





# Решаем задачу по учебнику

ассортимент

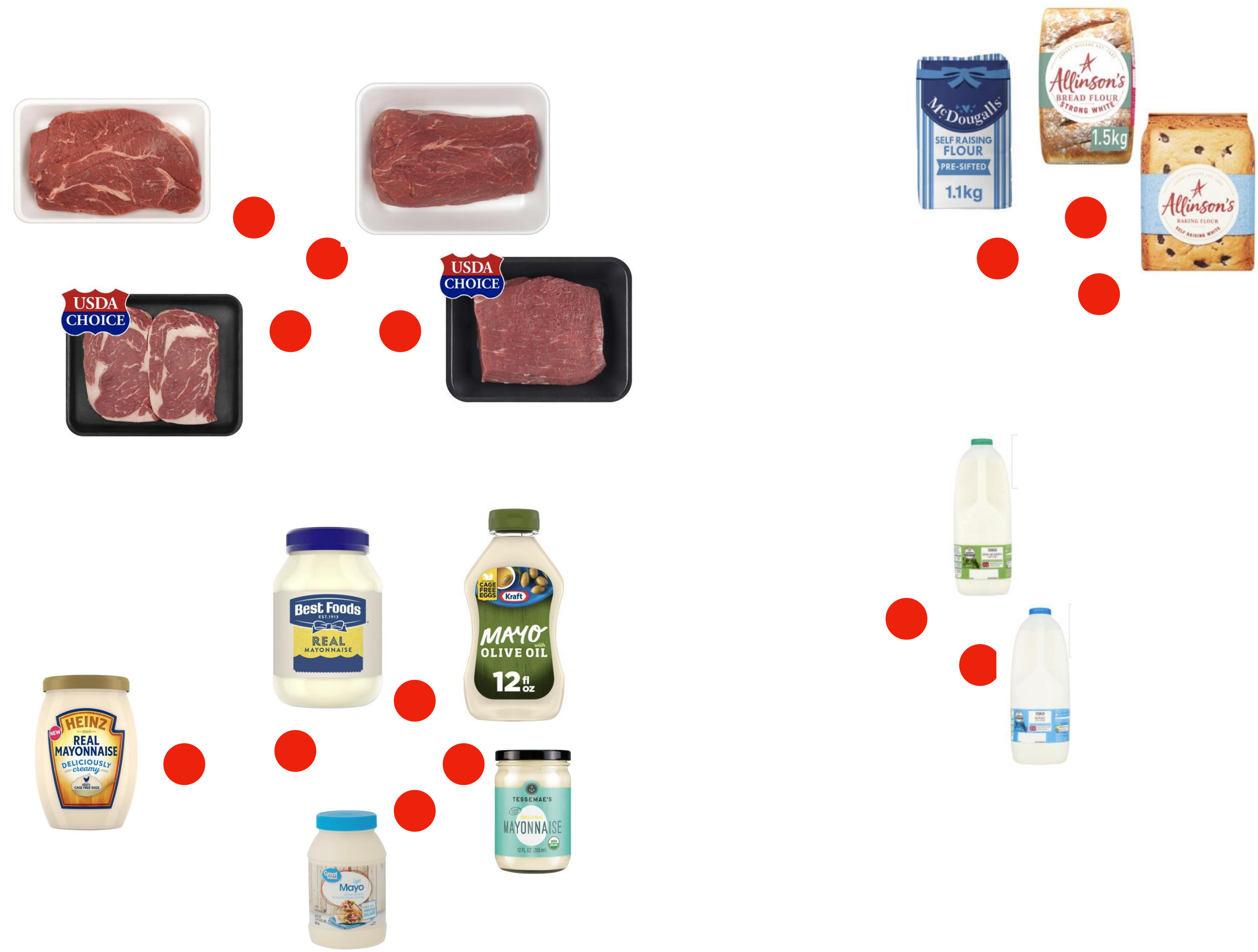
магазинов

## тексты ингредиентов

- 100g plain flour
- 2 large eggs
- 300ml milk



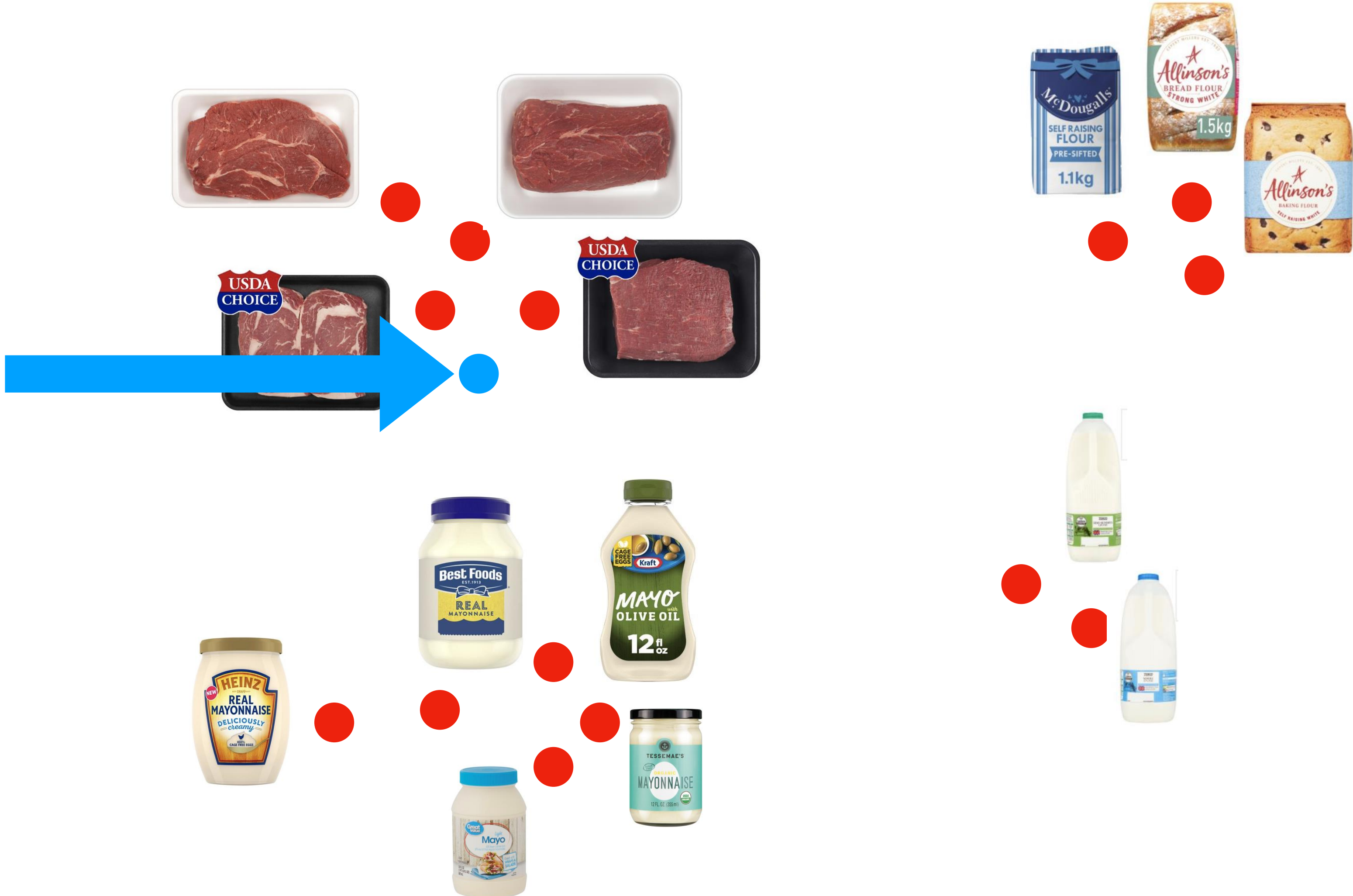
# Решаем задачу по учебнику





# Решаем задачу по учебнику

100g beef





# Решаем задачу по учебнику

100g beef



# Какая конкретно модель? Да неважно

## Хоть классика

$$w_{i,j} = tf_{i,j} \times \log \left( \frac{N}{df_i} \right)$$

$tf_{i,j}$  = number of occurrences of  $i$  in  $j$

$df_i$  = number of documents containing  $i$

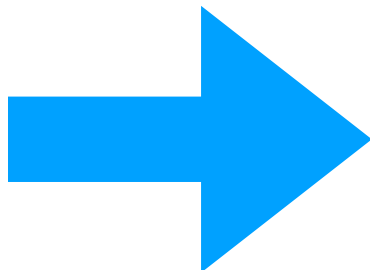
$N$  = total number of documents

## Хоть трансформеры



И пока аннотаторы трудятся...

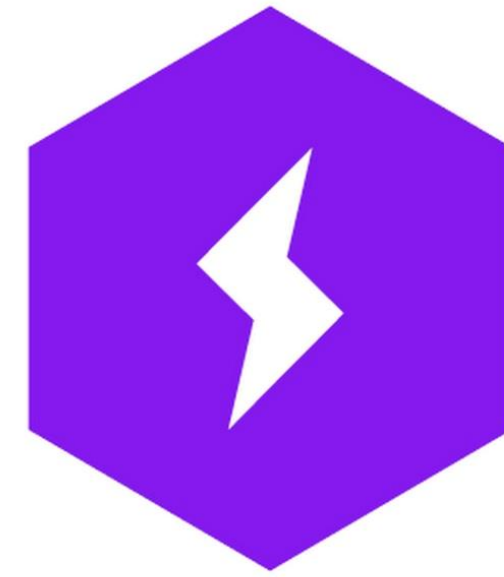
100g beef





ML-инженер отдыхает наслаждается автоматизацией

mlflow™



# Вариативность зашкаливает

225 grams of beef

# Вариативность зашкаливает

225 grams of beef



Tesco Finest Beef  
Roasting Joint



Tesco Beef Stir Fry Strips  
357G



Tesco Diced Beef 400G



Tesco 2 Beef Medallion  
Steaks 340G



# Вариативность зашкаливает

<num><unit>of beef



Tesco Finest Beef  
Roasting Joint



Tesco Beef Stir Fry Strips  
357G



Tesco Diced Beef 400G



Tesco 2 Beef Medallion  
Steaks 340G

Вариативность зашкаливает

<num><unit>of beef



Tesco Finest Beef  
Roasting Joint



Tesco Beef Stir Fry Strips  
357G



Tesco Diced Beef 400G



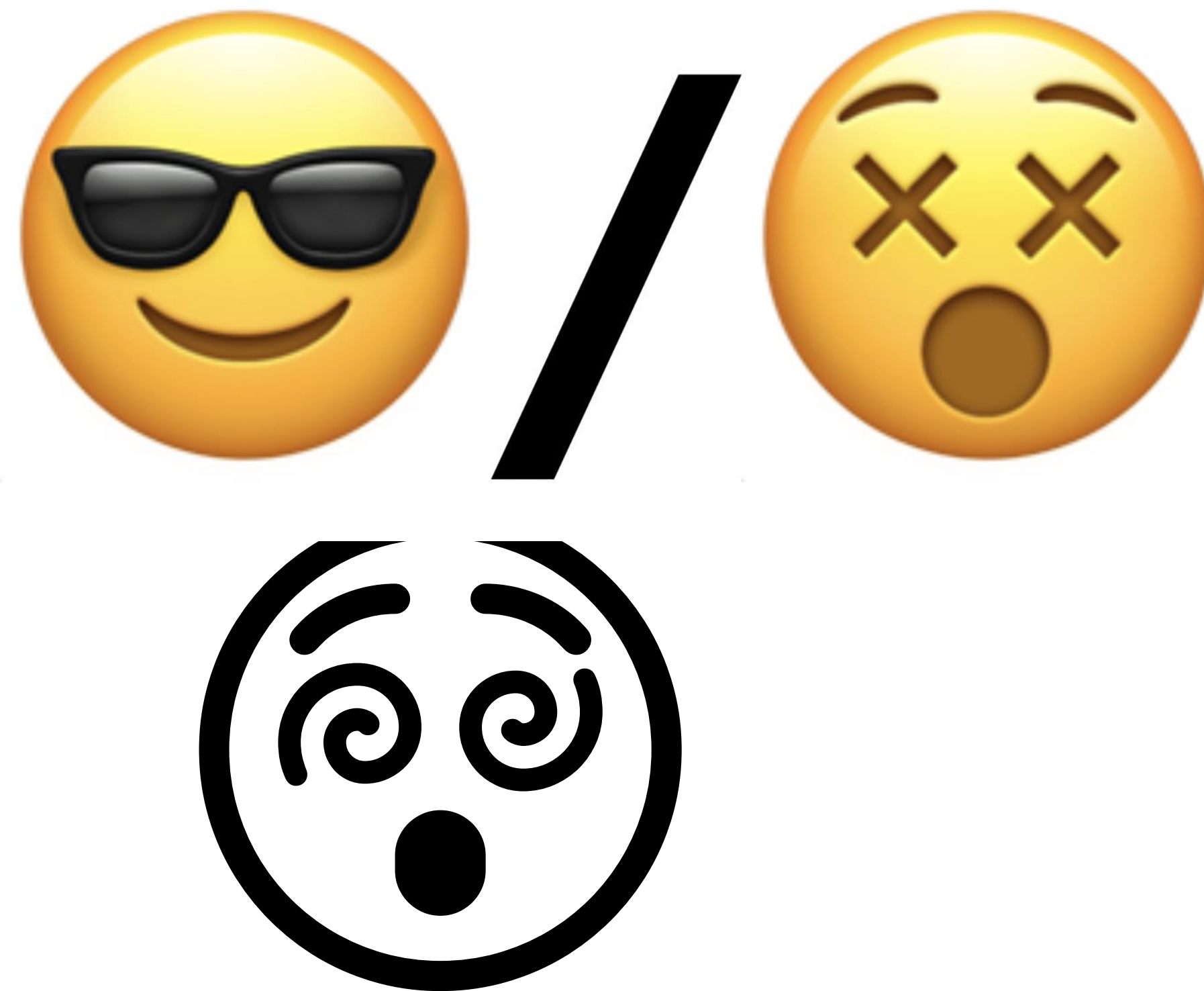
Tesco 2 Beef Medallion  
Steaks 340G

1lb **beef**, cut into bite-size pieces

2 cups diced cooked **beef**

2 lbs, **beef**, to serve four

ML умирает молча





А с точки зрения бизнеса?

Walmart 



# Показываем типичные метрики



CTR = 5%  
ACCURACY = 95%

# QA клиента

Walmart 



CTR = 5%  
ACCURACY = 95%



CTR = ??  
ACCURACY = 25%



Вторая метрика теперь тоже под вопросом

Walmart 



CTR = 5%  
ACCURACY = 95%



CTR = ??  
ACCURACY = 25%

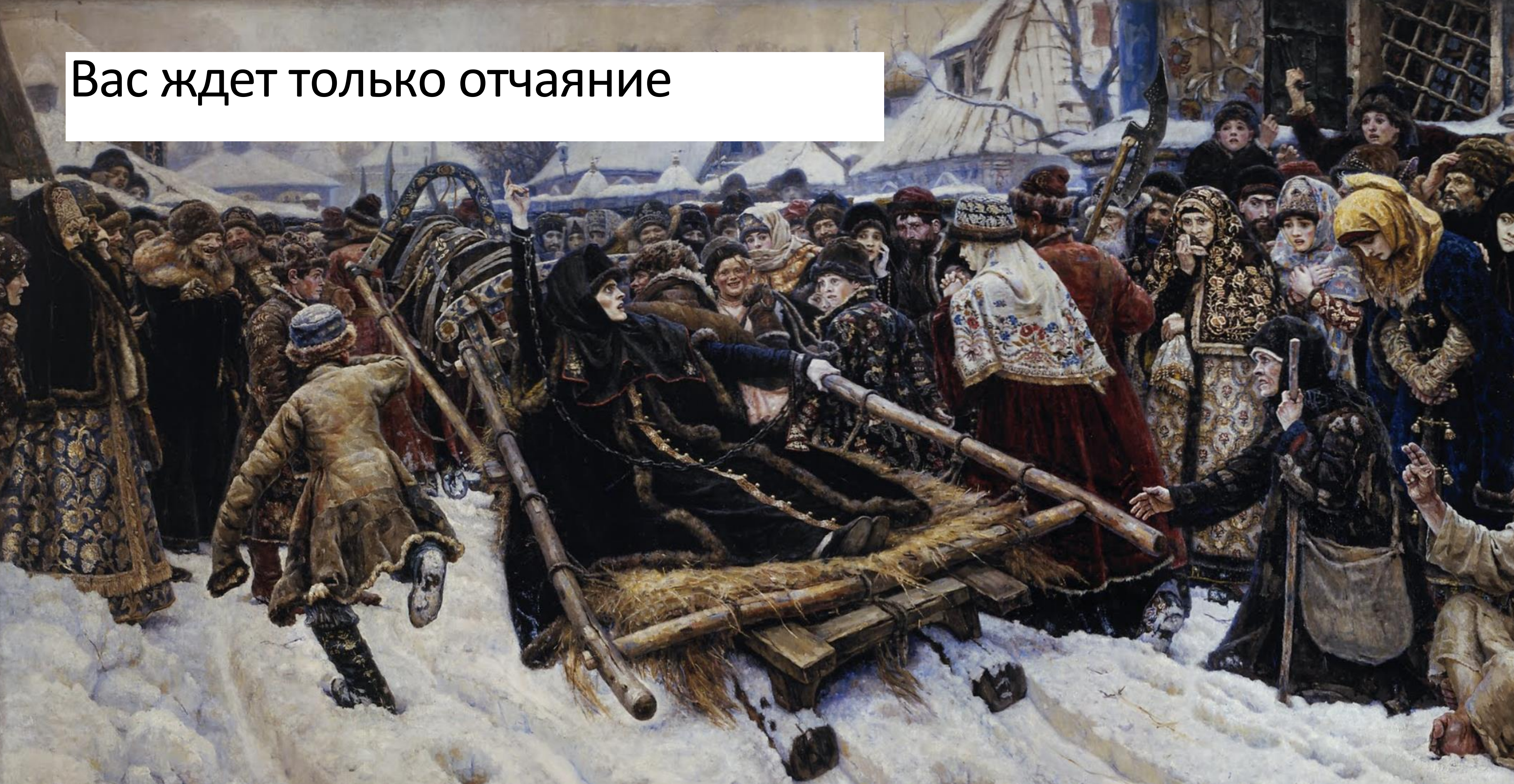
# А внутри команды?



“Мы стали проводить А/А/В вместо А/В, и оказалось, что мы занимались ерундой”



Вас ждет только отчаяние



Василий Иванович Суриков. ML-инженер покидает компанию, устав ругаться с продуктами



# Что реально нужно бизнесу

# Что реально нужно бизнесу

- Прозрачность

# Что реально нужно бизнесу

- Прозрачность
- Предсказуемость



# Что реально нужно бизнесу

- Прозрачность
- Предсказуемость
- Масштабируемость

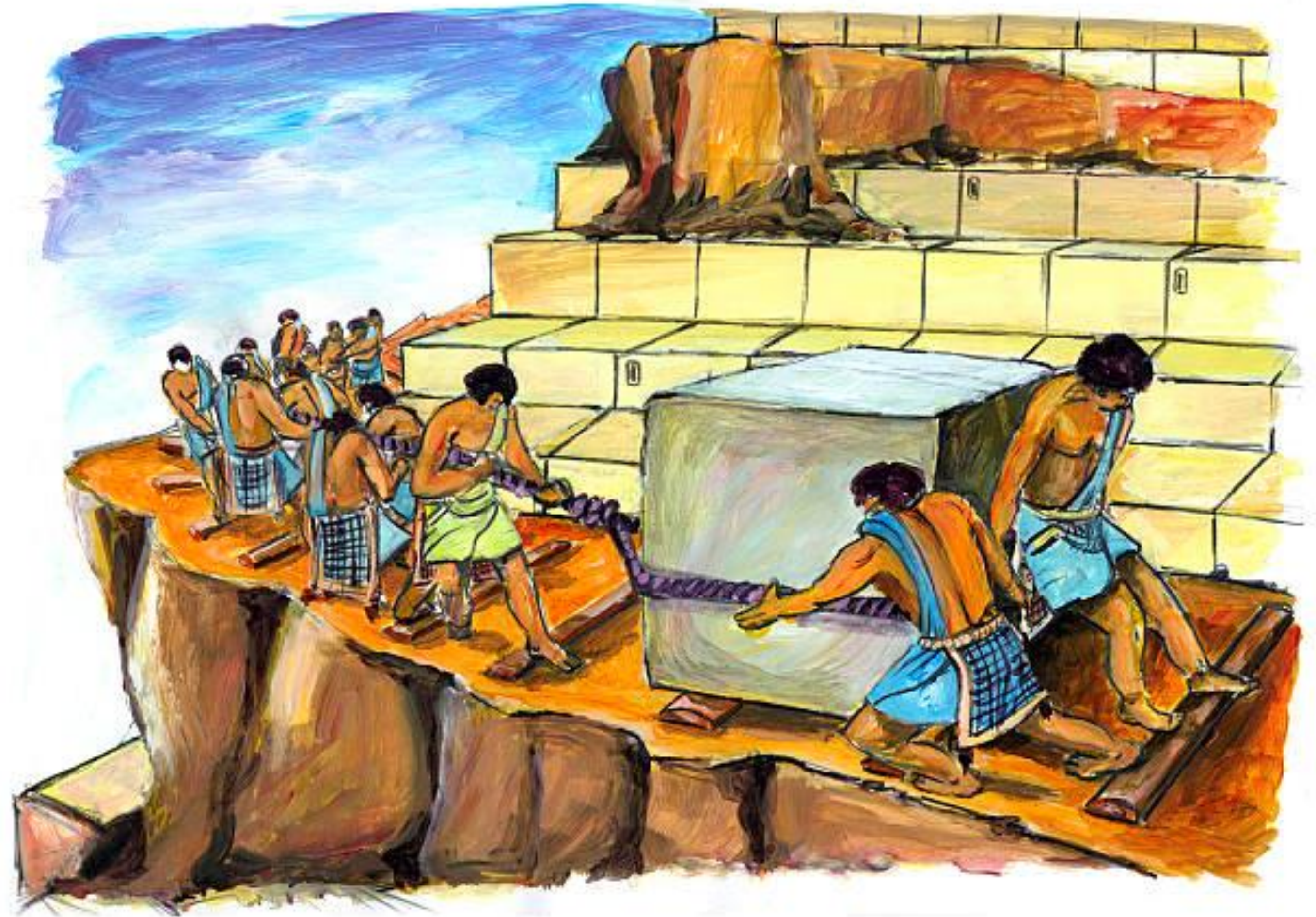


# Что реально нужно бизнесу

Прозрачность

Предсказуемость

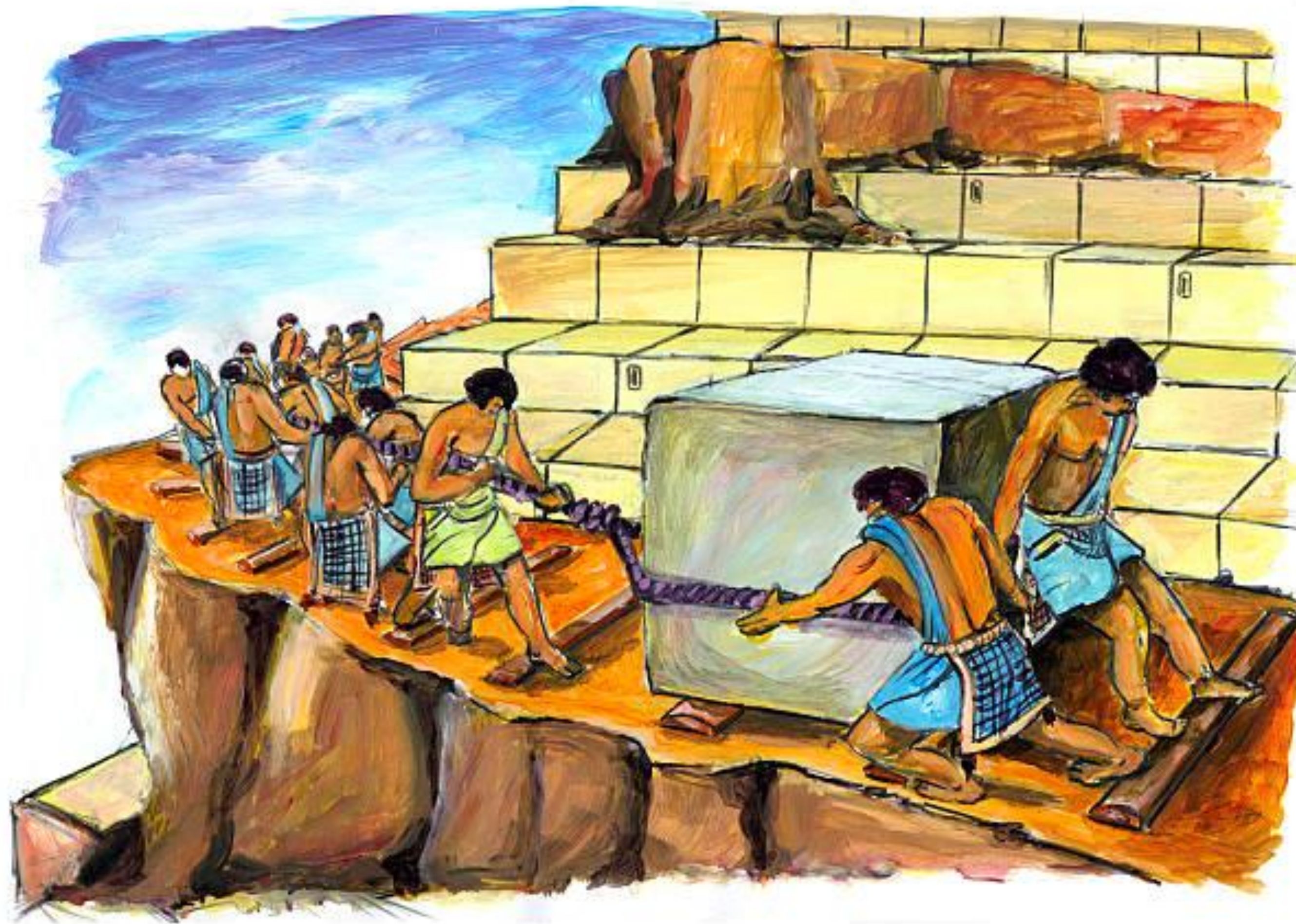
Масштабируемость





# Что реально нужно бизнесу

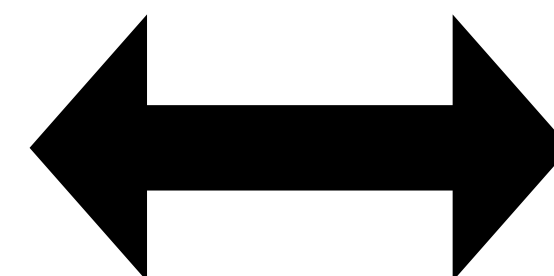
- ✓ Прозрачность
- ✓ Предсказуемость
- ✗ Масштабируемость



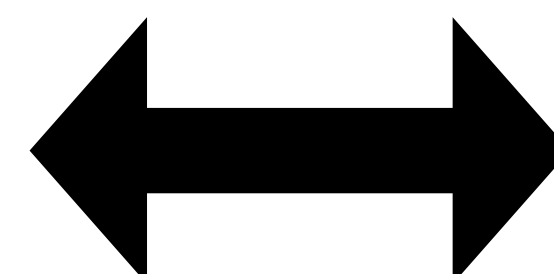


Похожие по смыслу строки

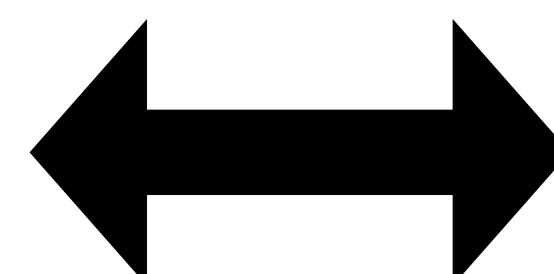
**225 grams of beef**



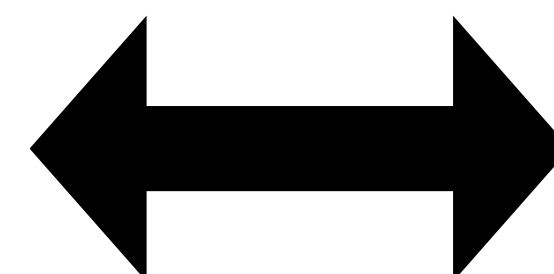
**1lb beef, cut into pieces**



**2 cups diced cooked beef**



**2 lbs, beef, to serve four**



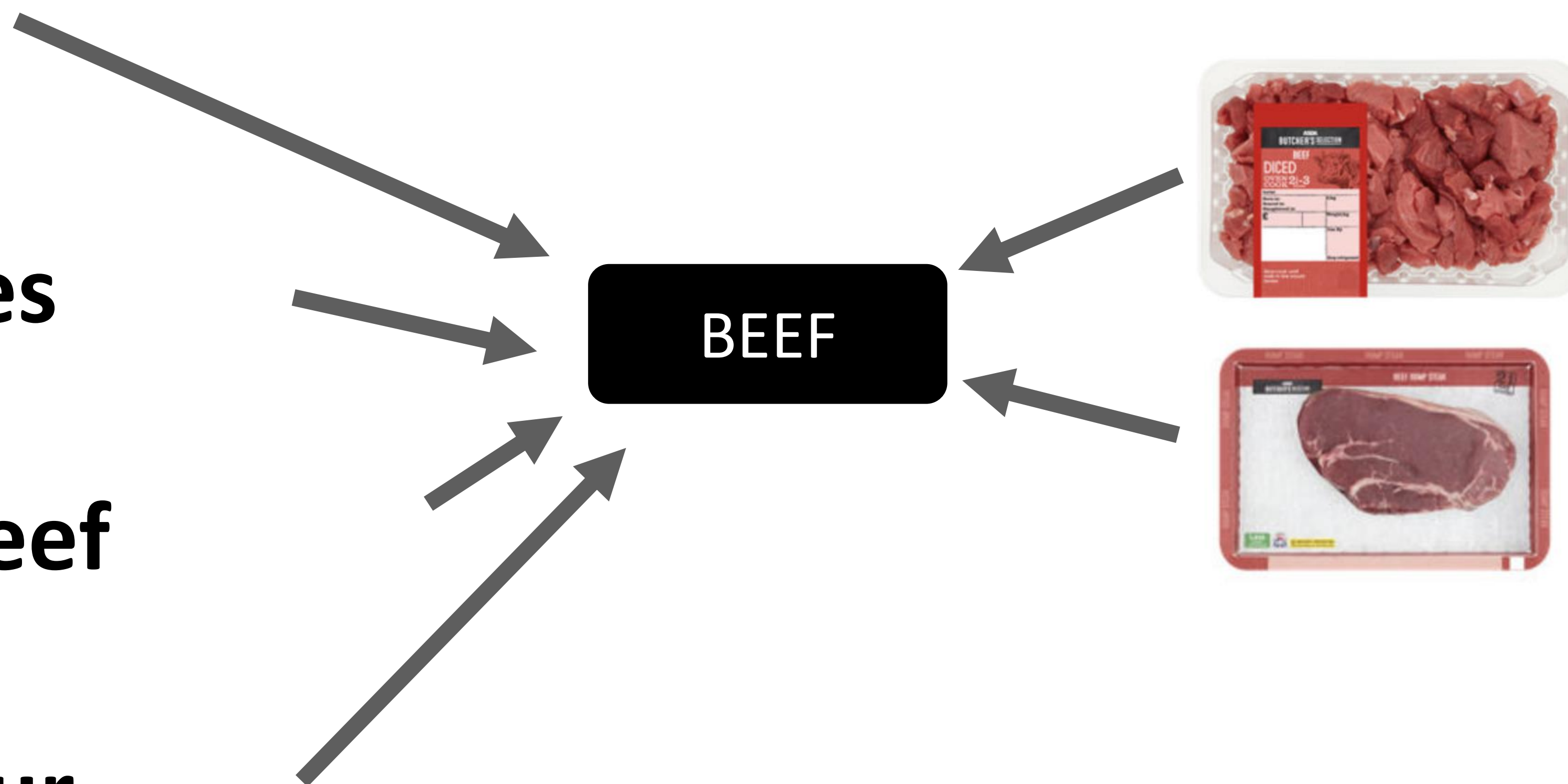
# Нормализованные сущности

**225 grams of beef**

**1lb beef, cut into pieces**

**2 cups diced cooked beef**

**2 lbs, beef, to serve four**





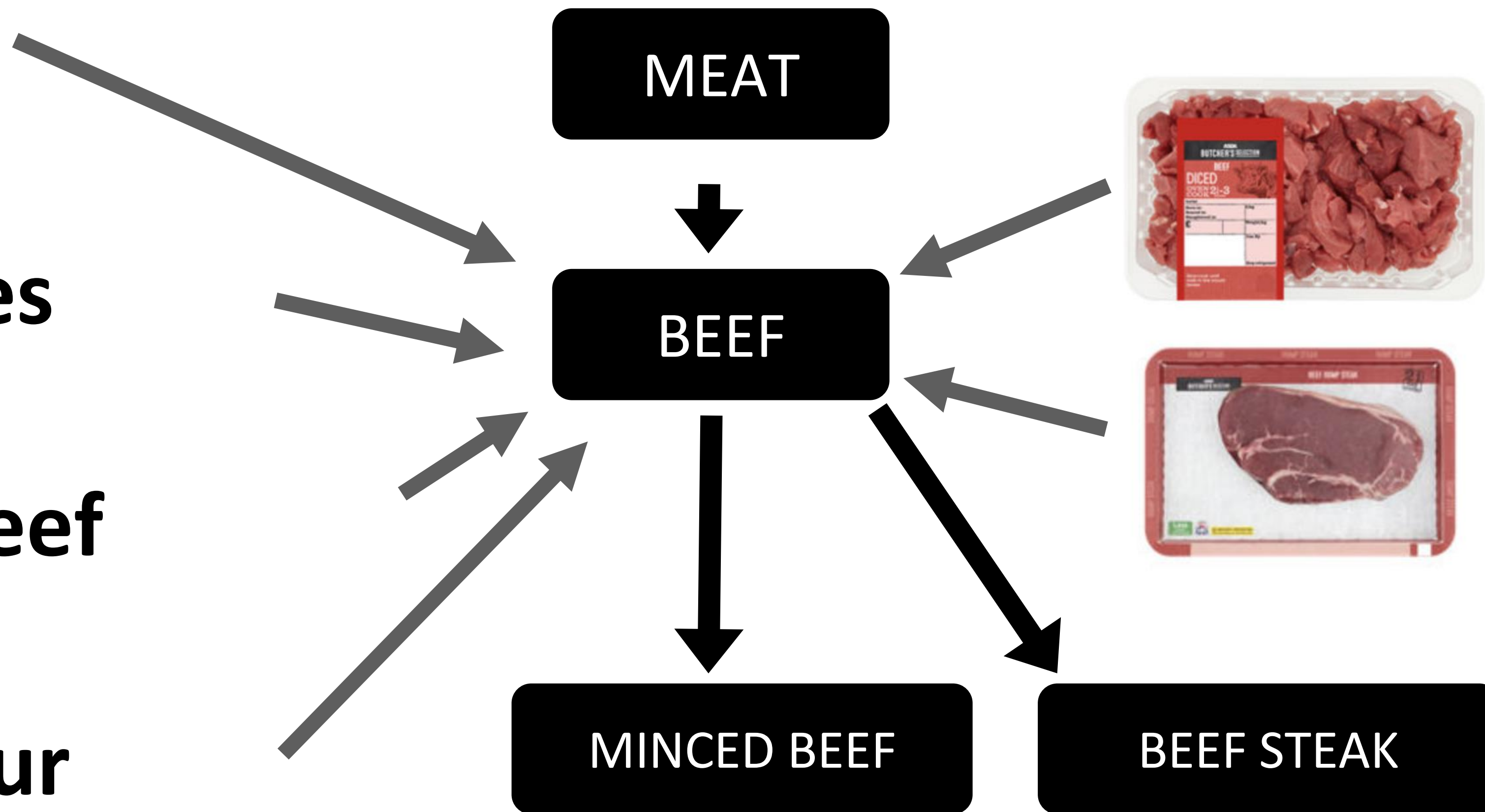
# Онтология сущностей

**225 grams of beef**

**1lb beef, cut into pieces**

**2 cups diced cooked beef**

**2 lbs, beef, to serve four**



# Как управляться с огромной онтологией?

**225 grams of beef**



be|

BEAPOT  
GUMMY BEAR  
BELUGA CAVIAR  
BEEF  
MINCE BEEF  
BEEF STEAK  
BEEF CUBE  
BEEF



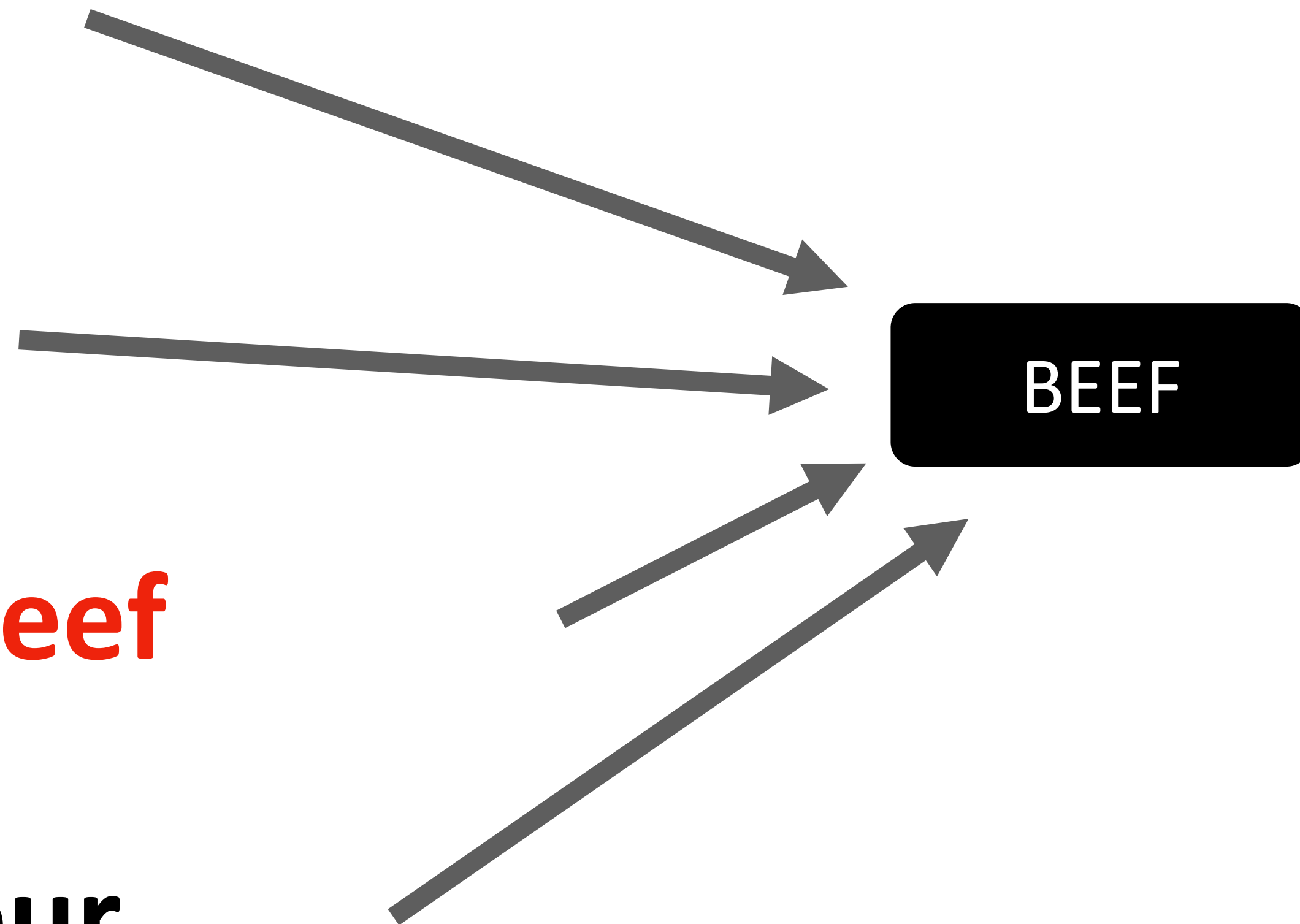
# Мәппинг через сабсринги

225 grams of **beef**

1lb **beef**, cut into...

2 cups diced cooked **beef**

2 lbs, **beef**, to serve four

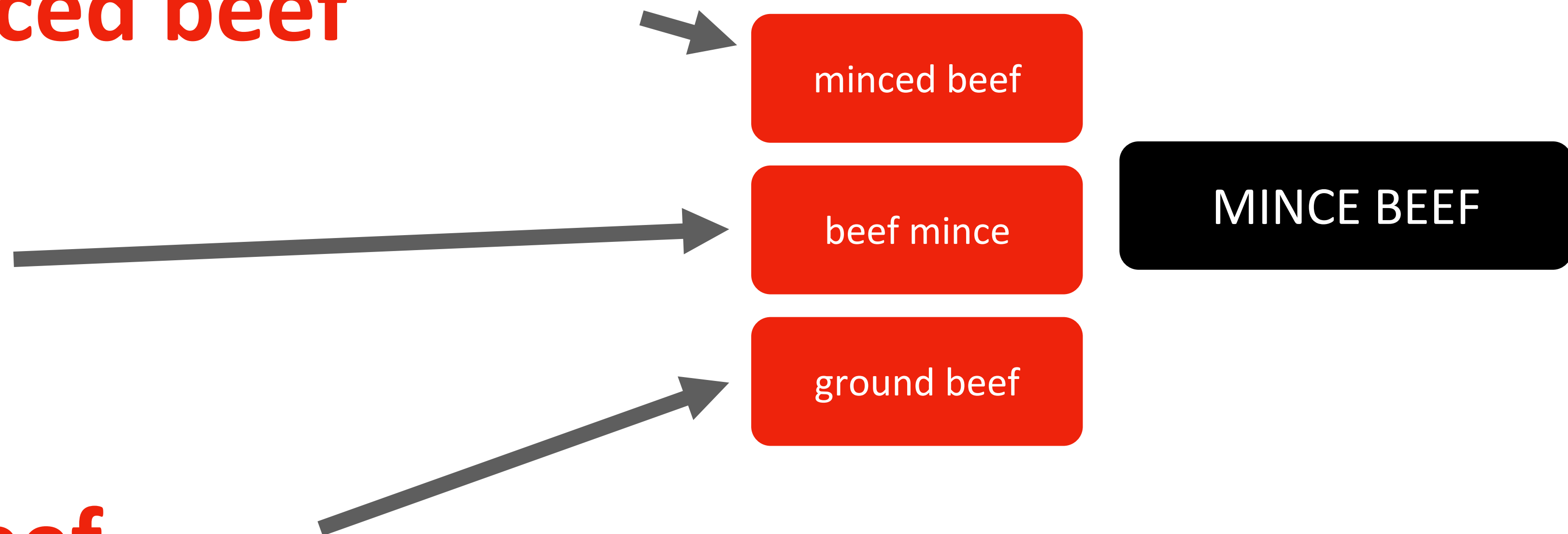


# Маппинг через сабсринги

225 grams of **minced beef**

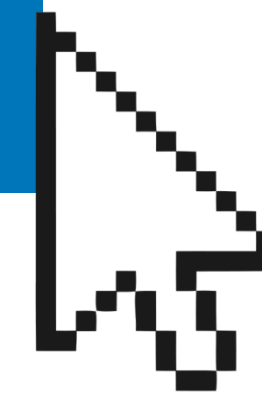
1lb **beef mince**

2 lbs of **ground beef**

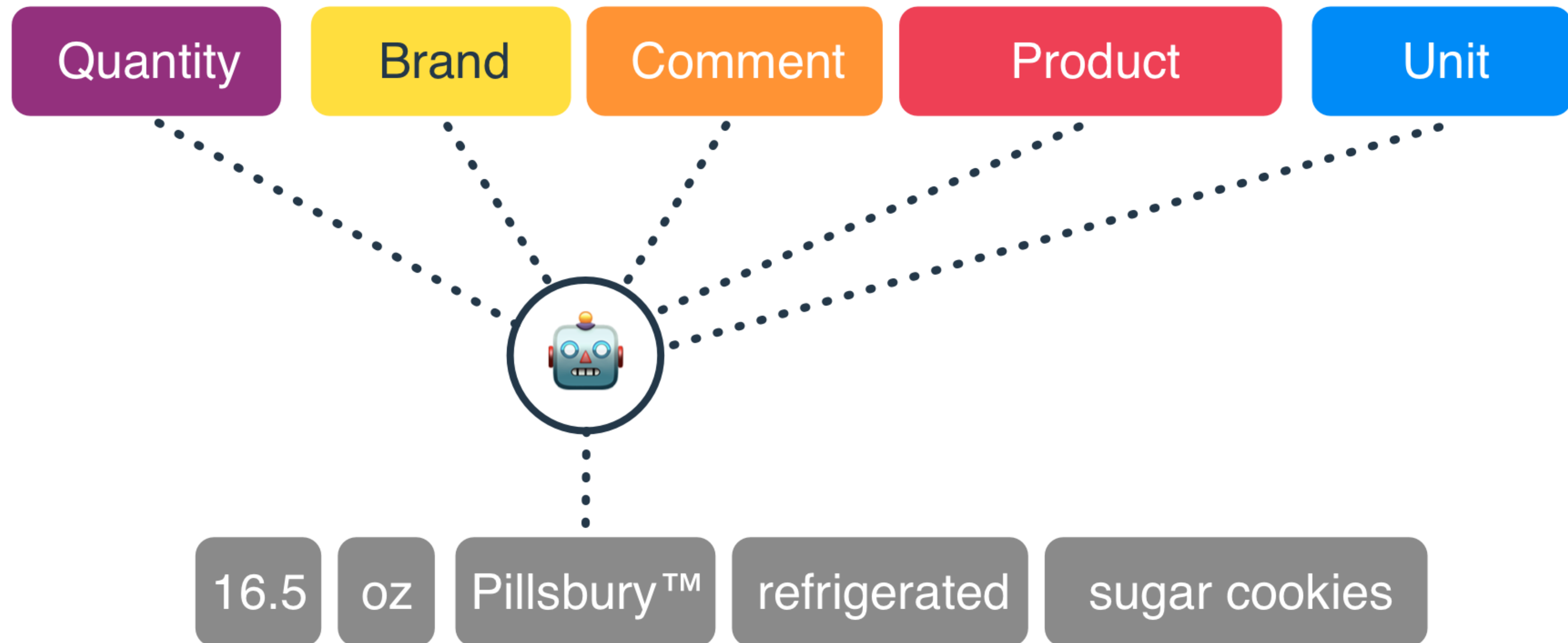


# Маппинг через сабсринги

225 grams of **minced beef**



# Задача Sequence tagging





# Задача Sequence tagging



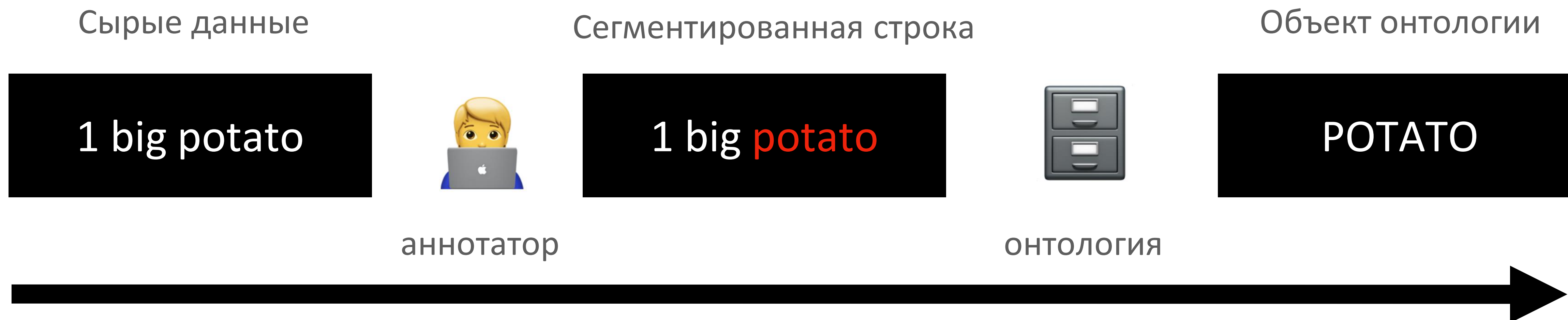
Quantity: 16.5

Unit: oz

Brand: Pillsbury

Product: sugar cookies

# Максимально эффективная ручная работа



# Здесь есть место для ML





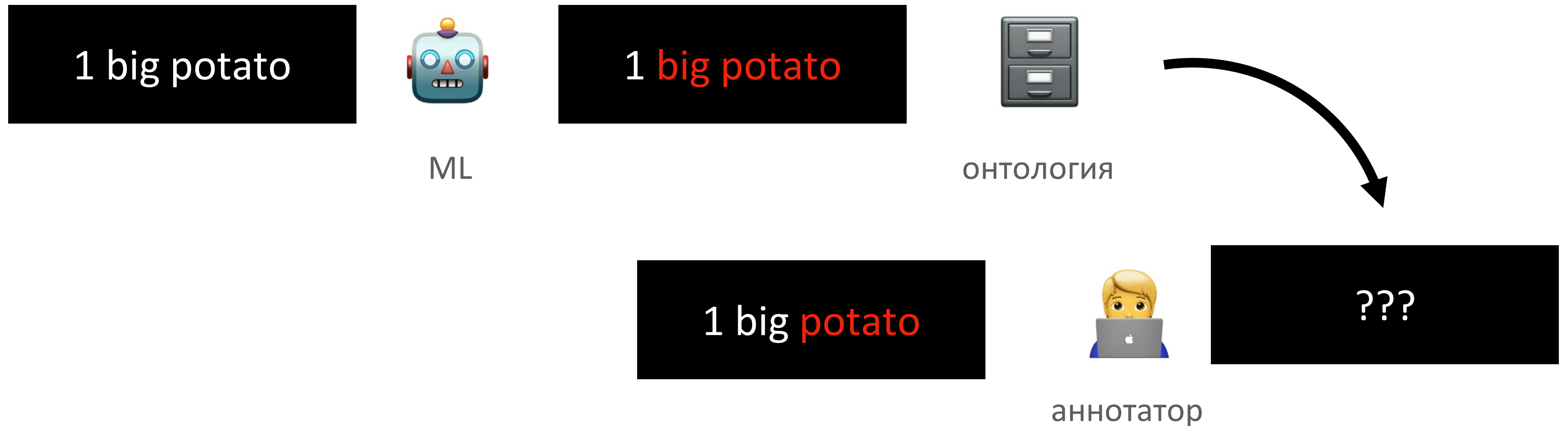
# Здесь есть место для ML



# Здесь есть место для ML



# Здесь есть место для ML

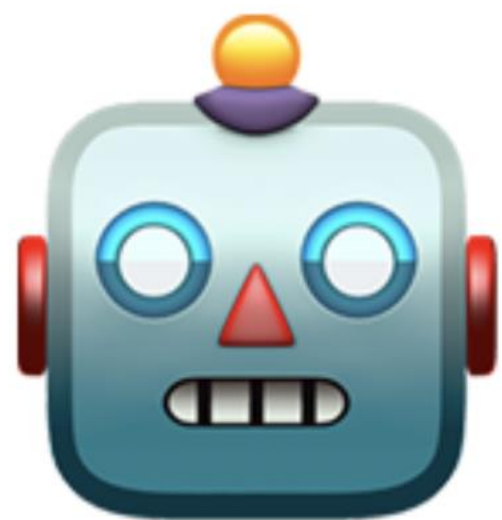




# Отказ от классификации!

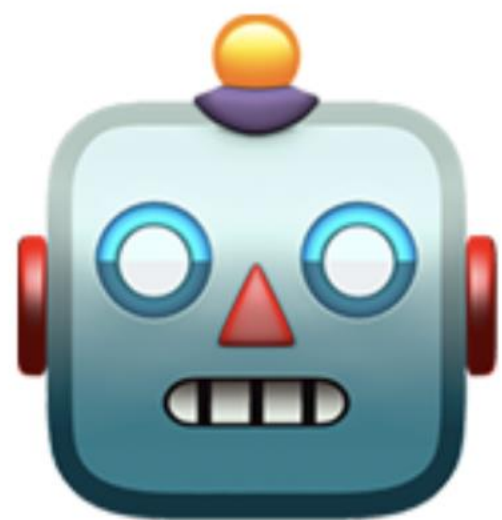


# Можно и в чистом ML-стапе, но сложно

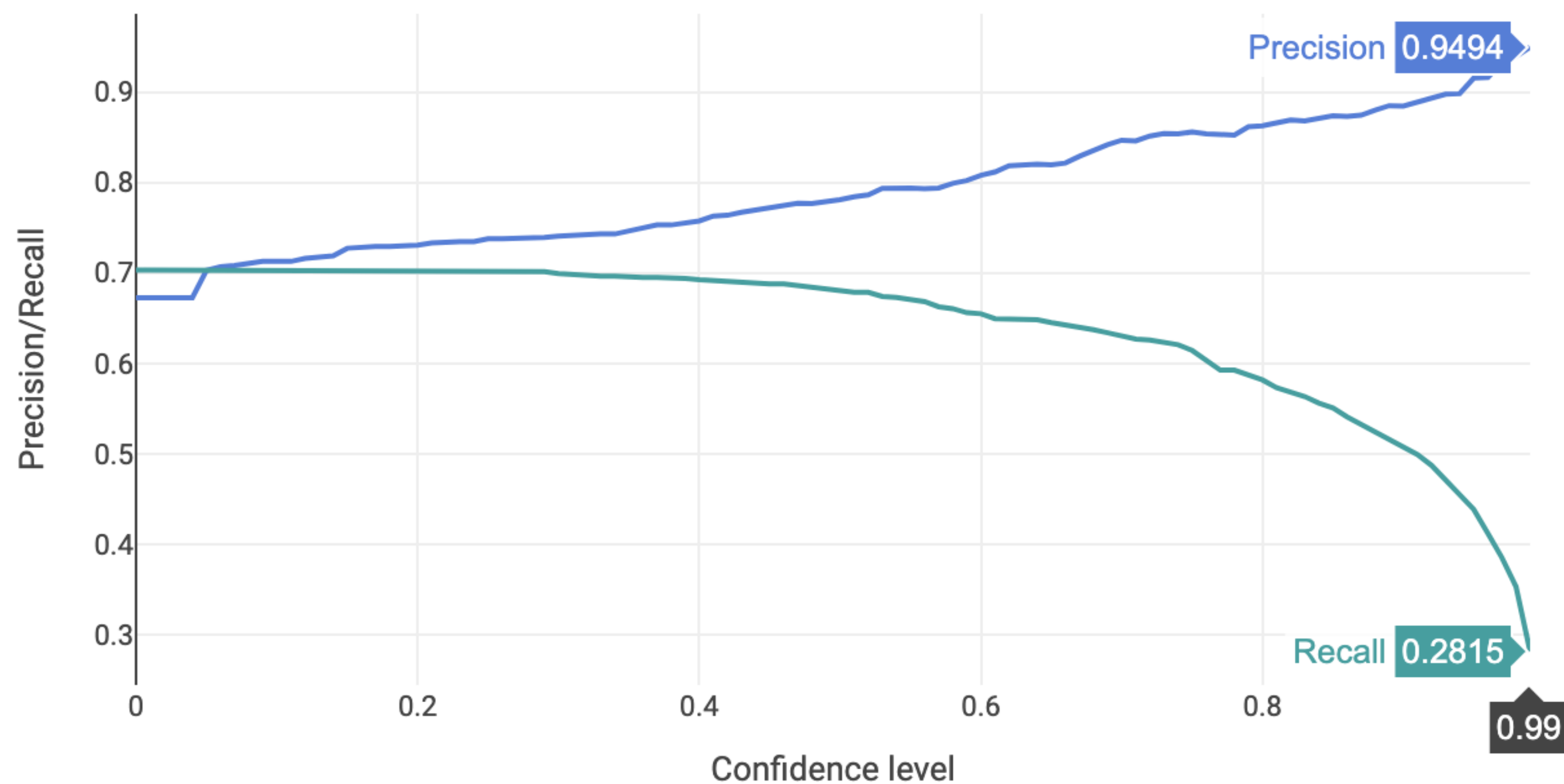


“это говядина с вероятностью 70%”

# Можно и в чистом ML-стапе, но сложно



“это говядина с вероятностью 70%”



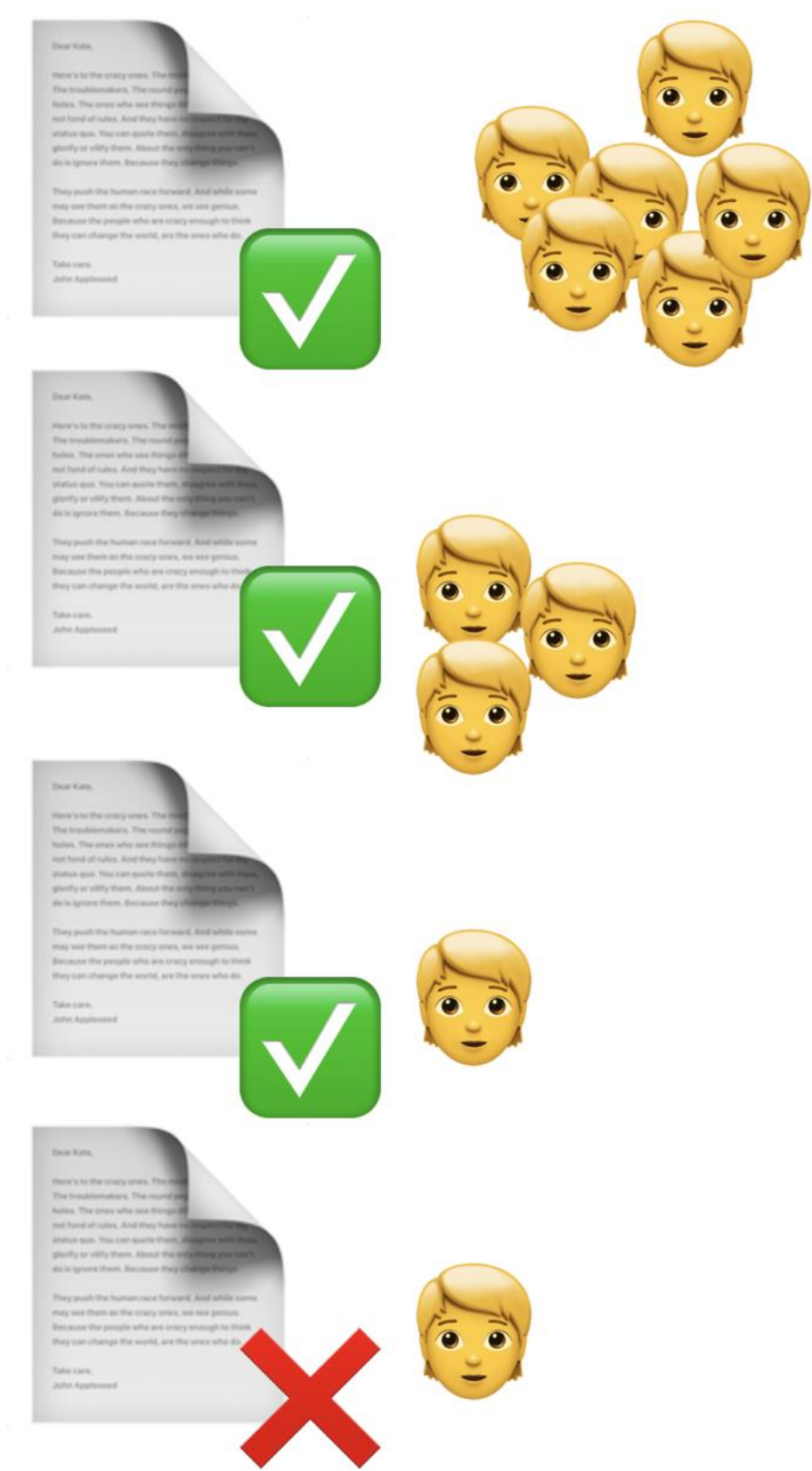


# Появляются качественные офлайн-метрики



75% are parsed successfully

# Появляются качественные офлайн-метрики



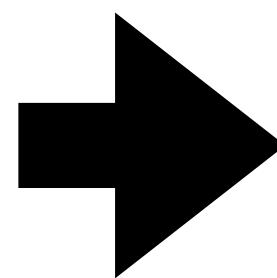
99% users are happy

# Бизнес-процесс



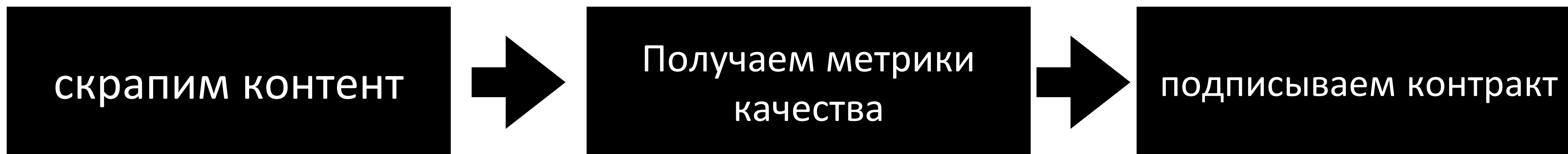
# Бизнес-процесс

скрапим контент



Получаем метрики  
качества

# Бизнес-процесс

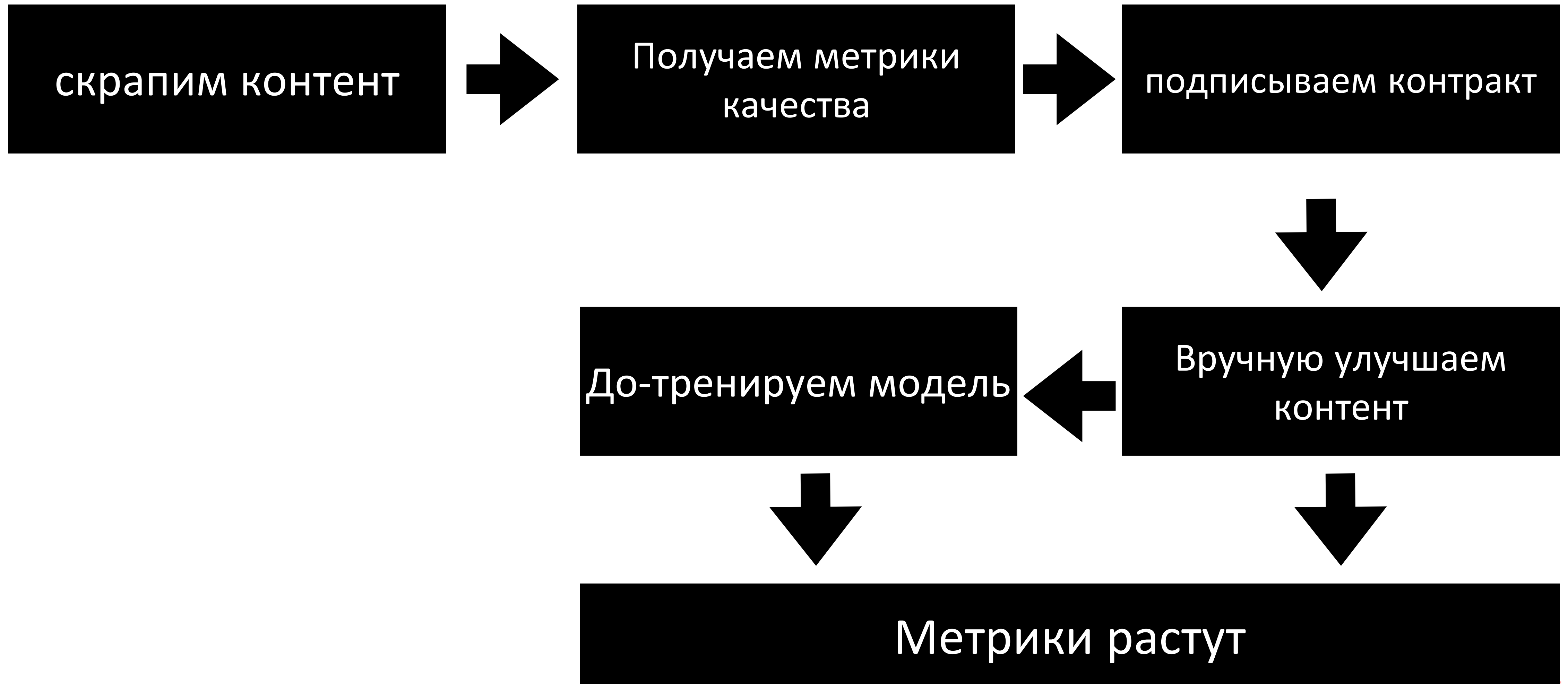


# Бизнес-процесс

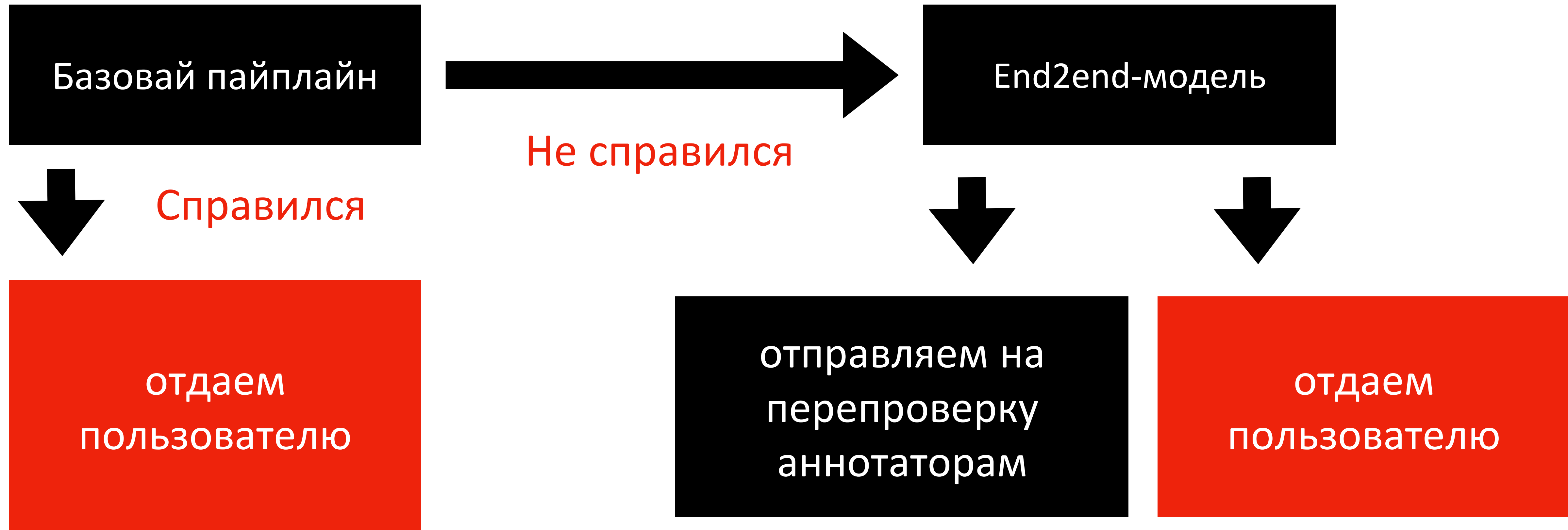




# Бизнес-процесс



# А потом можно сделать совсем хорошо



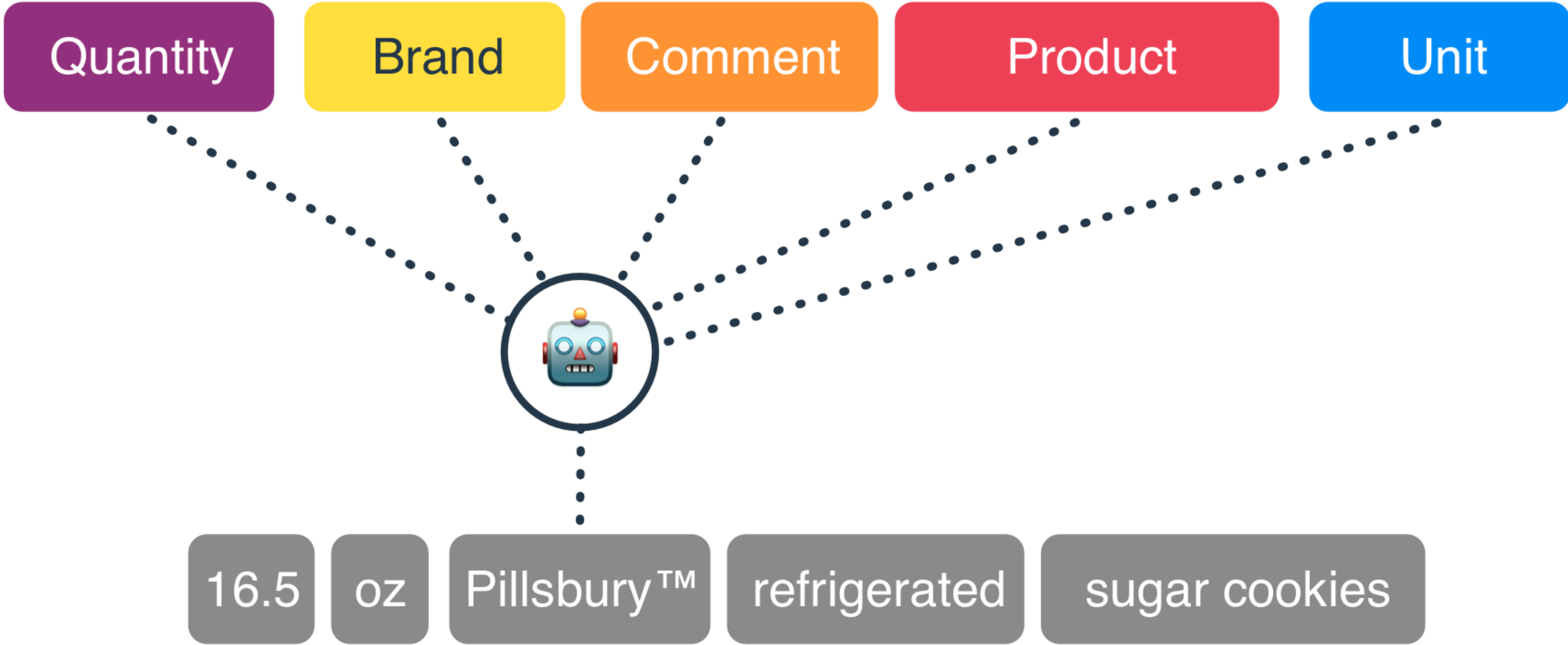
Резюмируя





# Bonus round! Тонкости разметки

# Bonus round! Тонкости разметки



# Bonus round! Тонкости разметки



Quantity: 16.5

Unit: oz

Brand: Pillsbury

Product: sugar cookies



# Bonus round! Тонкости разметки

2 ( 12 ounces ) cans black beans

---

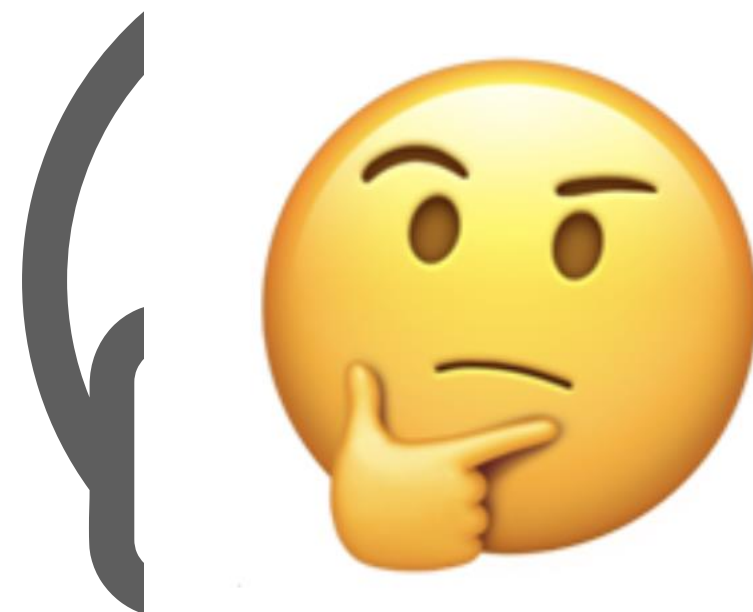
??

# Что делать?

2 ( 12 ounces ) cans black beans

---

??

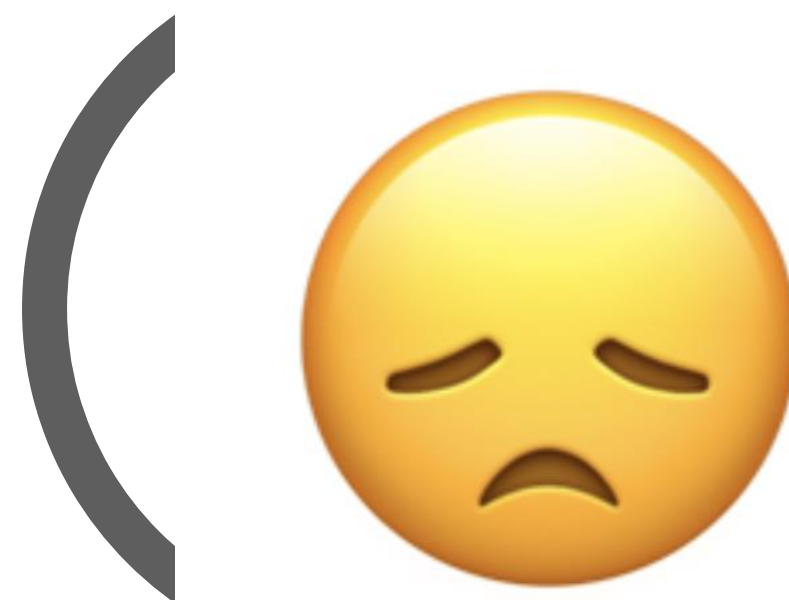


Как нам помочь аннотатору сделать правильный выбор?

2 ( 12 ounces ) cans black beans

---

??

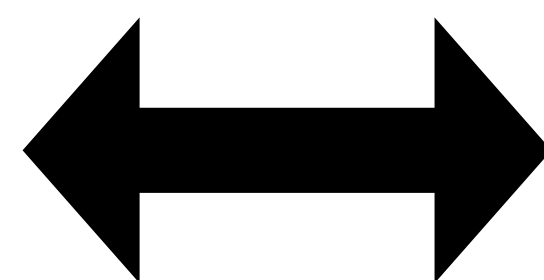




# Как нам помочь аннотатору сделать правильный выбор?

2 (15 ounce) cans  
black beans, rinsed  
and drained

and drained

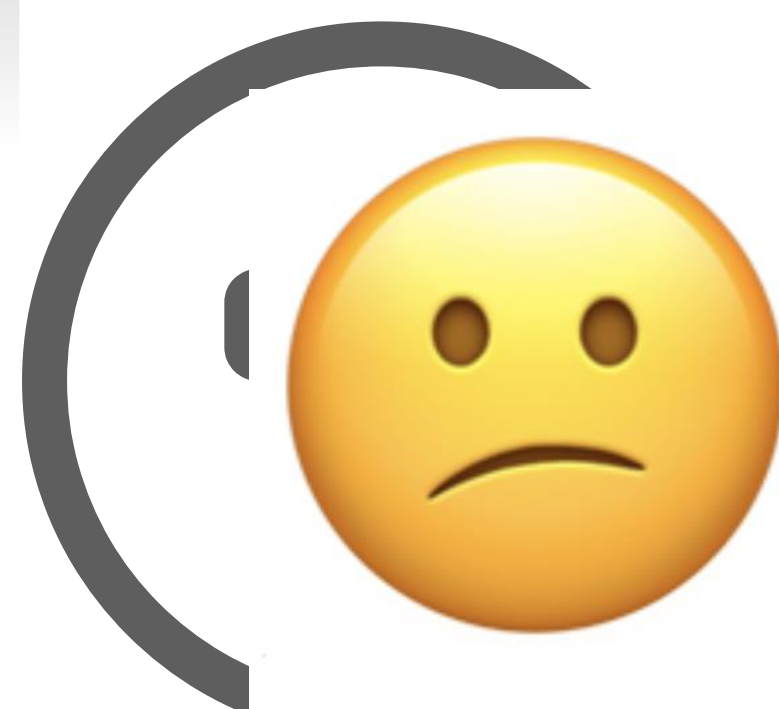


Quantity: 30

Unit: oz

Brand: -

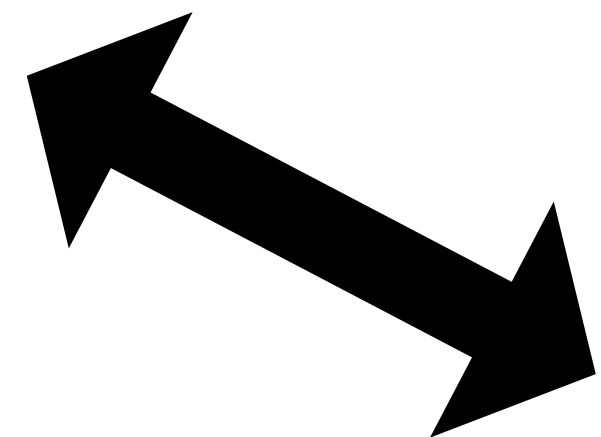
Product: black beans



# Коммуникации с инжинирингом



Инженеры



Аннотаторы

(трудности построения онтологий)



ORGANIC ABERDEEN ANGUS BEEF RUMP ROAST

# (трудности построения онтологий)



## ORGANIC ABERDEEN ANGUS BEEF RUMP ROAST

### Frequent questions and answers

*Q: Is nutrition important (no salt added, reduced sugar, sugar free...)?*

*A: Yes, as noted in the example above we should separate those products from regular ones.*

*Q: Why?*

*A: Because of nutrition, it is different, but also store items matching. Important: low-fat, reduced-fat, and fat-free are not synonyms, they all are separated products.*

---

*Q: Singular or plural?*

*A: Prioritize singular except in cases when plural sounds more 'natural'.*

*Example: CANNED OLIVES should be in plural because there is not only one olive in a can.*

---

*Q: Is gluten-free, dairy-free, vegan comment or part of the product name?*

*A: Depending on the context, only when it makes sense. if the product is GLUTEN FREE BREAD it's a part of the product name because BREAD usually contains gluten. But in cases like gluten-free milk it's a comment because milk doesn't contain gluten.*

---



# Проблемы данных тяжело править постфактум

данные > ML-алгоритмы

Сухой остаток



Сухой остаток





